

Catch'Em All: Locating Multiple Diffusion Sources in Networks with Partial Observations

Kai Zhu* ¹, Zhen Chen^{†2} and Lei Ying^{‡2}

¹Google Inc.

²School of of Electrical, Computer and Energy Engineering, Arizona State University

Abstract

This paper studies the problem of locating multiple diffusion sources in networks with partial observations. We propose a new source localization algorithm, named Optimal-Jordan-Cover (OJC). The algorithm first extracts a subgraph using a candidate selection algorithm that selects source candidates based on the number of observed infected nodes in their neighborhoods. Then, in the extracted subgraph, OJC finds a set of nodes that “cover” all observed infected nodes with the minimum radius. The set of nodes is called the Jordan cover, and is regarded as the set of diffusion sources. Considering the heterogeneous susceptible-infected-recovered (SIR) diffusion in the Erdős-Rényi (ER) random graph, we prove that OJC can locate all sources with probability one asymptotically with partial observations. OJC is a polynomial-time algorithm in terms of network size. However, the computational complexity increases exponentially in m , the number of sources. We further propose a low-complexity heuristic based on the K-Means for approximating the Jordan cover, named Approximate-Jordan-Cover (AJC). Simulations on random graphs and real networks demonstrate that both AJC and OJC significantly outperform other heuristic algorithms.

1 Introduction

Diffusion source localization (or called the information source detection) is to infer the source(s) of an epidemic diffusion process in a network based on some observations of the diffusion. Possible observed information includes node states (e.g., infected or susceptible) and the timestamps at which nodes changed their states. The solution to this problem has a wide range of applications. In epidemiology, identifying patient zero helps diagnose the cause and the origin of the disease. For cybersecurity, tracing the source of malware is an important step in the investigation of a cyber attack. On online social networks, the trustworthiness of news/information heavily depends on its source.

Since the seminal work of Shah and Zaman [15], the problem has received a lot of attention from different research communities, including machine learning, signal processing and information theory (a detailed discussion can be found in the related work). However, most existing results assume that the diffusion is from single source and the observed information is one or multiple complete network snapshot(s). Furthermore, provably guarantees on the detection rate have only been established for tree networks except

*zhukai93@gmail.com

†Zhen.Chen.1@asu.edu

‡Lei.Ying.2@asu.edu

a recent paper by Zhu and Ying [24], where they proposed a Short-Fat Tree (SFT) algorithm and proved that under the single-source, homogeneous Independent-Cascade (IC) model, SFT locates the actual source in the Erdős-Rényi random graph [4] with probability one asymptotically (as the size of the network increases) when a complete snapshot of the network is given.

In this paper, we consider the diffusion source localization problem in a setting that generalizes the existing ones at several important directions.

- *Multiple sources versus single source:* In this paper, the diffusion can be originated from multiple nodes simultaneously, instead of from a single source. When the infection duration is sufficiently short, the infected subnetworks from different sources are disconnected components. In such cases, the single-source localization algorithms can be applied to each of the infected subnetwork. We, however, do not make such an assumption, and consider the scenario where the infected subnetworks may overlap with each other, so the single-source localization algorithms cannot be directly applied.
- *A partial snapshot versus a complete snapshot:* In this paper, we assume a partial snapshot in which each node reports its state with some probability, which is in contrast to a complete snapshot assumed in the literature where all nodes' states are observed. Because of a partial snapshot, the sources may not report their states and be observed as infected nodes; and the observed infected nodes may not form a connected component. Both increase the uncertainty and complexity of the problem. In fact, it turns out to be critical to have a candidate selection algorithm to select source candidates from unobserved nodes but only use observed infected nodes in computing the infection eccentricity. The selection step yields $27\times$ reduction on the computing time in our simulations while guaranteeing the same the detection rate, and yields $600\times$ reduction on the computing time with a slight reduction of the detection rate.
- *Heterogeneous diffusion versus homogeneous diffusion:* Our algorithm applies to the heterogeneous SIR diffusion model where links have different infection probabilities and nodes have different recovery probabilities. The asymptotic guarantees on the detection rate hold for the heterogeneous SIR model.

While some of these extensions have been investigated in the literature individually, our model includes all three extensions. We propose a novel algorithm for locating multiple sources for such a general model and prove theoretical guarantees on the detection rate for non-tree networks. The main results of the paper are summarized below.

- (1) We introduce the concept of Jordan cover, which is an extension of Jordan center. Loosely speaking, a Jordan cover with size m is a set of m nodes that can reach all *observed* infected nodes with the minimum hop-distance. We propose Optimal-Jordan-Cover (OJC), which consists of two steps: OJC first selects a subset of nodes as the set of the candidates of the diffusion sources; and then it finds a Jordan cover in the subgraph induced by the candidate nodes and the observed infected nodes. We emphasize that only the hop-distance to the observed infected nodes is considered in computing a Jordan cover.
- (2) We analyze the performance of OJC on the ER random graph, and establish the following performance guarantees.
 - (i) When the infection duration is shorter than $\frac{2}{3} \frac{\log n}{\mu}$, where μ is the average node degree and n is the number of nodes in the network, OJC identifies the sources with probability one asymptotically as n increases.

- (ii) When the infection duration is at least $\left\lceil \frac{\log n}{\log \mu + \log q} \right\rceil + 2$ where q is the minimum infection probability, *under any source location algorithm*, the detection rate diminishes to zero as n increases under the Susceptible-Infected (SI) and Independent-Cascade (IC) models, which are special cases of the SIR model.
- (3) The computational complexity of OJC is polynomial in n , but exponential in m . We further propose a heuristic based on the K-Means for approximating the Jordan cover, named Approximate-Jordan-Cover (AJC). Assuming a constant number of iterations when using the K-Means, the computational complexity of AJC is $O(nE)$, where E is the number of edges. Our simulations on random graphs and real networks demonstrate that both AJC and OJC significantly outperform other heuristic algorithms.

1.1 Related Work

Shah and Zaman [15, 16] studied the rumor source detection problem, and proposed a new graph centrality called *rumor centrality*. They proved that the node with the maximum rumor centrality is the maximum likelihood estimator (MLE) of the diffusion source on regular trees under the continuous-time SI model. In addition, it has been proved that the detection probability is greater than zero on regular trees and approaches one for geometric trees. [17] analyzed the detection probability of rumor centrality for general random trees. Later, the performance of rumor centrality has been studied under different models, including multiple sources [11], incomplete observations [8], multiple independent diffusion processes from the same source [18].

Kai and Ying [21] proposed a sample path based approach for single source detection under the SIR model and introduced the concept of *Jordan infection center*. Assuming the homogeneous SIR model, a complete network snapshot and regular tree networks, they [21] proved that the Jordan infection center is the root of the most likely diffusion sample path, and it is within a constant hop-distance from the actual source with a high probability. *Assuming tree networks*, the performance of the Jordan infection center has been further studied, including partial observations under the heterogeneous SIR model [22], multiple sources [3], the SI model and SIS model [12].

Besides the rumor centrality and sample path based approach, diffusion source localization has also been investigated from other perspectives: 1) [9] proposed a dynamical programming algorithm for the IC model; 2) A variation of eigen centrality was proposed in [14] to detect multiple sources under the SI model; 3) [10] derived the mean field approximation of the MLE of the actual source and proposed a dynamic message passing algorithm based on that. Furthermore, [13, 25, 1, 20, 5] have proposed algorithms to identify the source with partial timestamps.

In this paper, we propose two new source localization algorithms, OJC and AJC, based on the Jordan cover. OJC guarantees probability one detection asymptotically on the ER random graph under the multi-source, heterogeneous SIR model, and with an incomplete snapshot, which is the first multi-source localization algorithm with provable guarantees for non-tree networks. AJC is a low-complexity approximation of OJC.

2 Problem Formulation

We assume the network is represented by an undirected graph g . Denote by $\mathcal{E}(g)$ the set of edges and $\mathcal{V}(g)$ the set of nodes in graph g . Let n denote the number of nodes and E denote the number of edges. We further assume a heterogeneous SIR model for diffusion. In this model, each node has three possible states:

susceptible (S), infected (I) and recovered (R). Time is slotted. At the beginning of each time slot, each infected node (say node u) attempts to infect its neighbor (say node v) with probability q_{uv} , independently across edges. We call q_{uv} the *infection probability* of edge (u, v) . At the end of each time slot, each infected node (say node u) recovers with probability r_u , independent of other infected nodes. We call r_u the *recovery probability*. We further assume $q_{uv} \in (0, 1]$ for all edges $(u, v) \in \mathcal{E}(g)$ and $r_v \in [0, 1]$ for all nodes $v \in \mathcal{V}(g)$.

Note that the SIR model includes two important special cases. When the recovery probability is zero, the SIR model becomes the Susceptible-Infected (SI) model [2], where infected nodes cannot recover. When the recovery probability is one, the SIR model becomes the Independent Cascade (IC) model [6] by regarding both infected nodes and recovered nodes as active nodes and regarding the susceptible nodes as inactive nodes.

We assume the epidemic diffusion starts from m sources in the network. In other words, at time slot 0, m nodes (sources) are in the infected state and all other nodes are in the susceptible state. Denote by s_1, s_2, \dots, s_m the sources and \mathcal{S} the set of sources, i.e., $\mathcal{S} = \{s_1, s_2, \dots, s_m\}$. We assume m is a *constant independent of n* .

Finally, we assume that a *partial* snapshot of the network state at time slot t is given, with an unknown observation time t . In the snapshot, each infected or recovered node reports its state with probability $\theta_v \in (0, 1)$, independent of other nodes. If a node reports its state, we call it an observed node. Denote by \mathcal{I}' the set of observed infected and recovered nodes. In this paper, we call infected nodes and recovered nodes as “infected nodes” unless explicitly clarified.

Based on \mathcal{I}' , the source localization problem is to find \mathcal{S} that solves the following maximum likelihood (ML) problem

$$\mathcal{W}^* = \operatorname{argmax}_{\mathcal{W} \subset \mathcal{V}(g)} \Pr(\mathcal{S} = \mathcal{W} | \mathcal{I}').$$

Even with a single diffusion source, this problem is known to be a difficult problem [15, 21] on non-tree networks. Therefore, instead of solving the ML problem above, we are interested in algorithms with *asymptotic perfect detection*, i.e., finding all sources with probability one as the network size increases. We believe this alternative metric is reasonable because we often need to solve the problem for large-size networks such as online social networks, and an algorithm with asymptotic perfect detection can detect sources with a high probability when the network size is large.

3 Algorithms

In this section, we present OJC and AJC based on the concept of Jordan cover.

Define the hop-distance between a node v and a *node set* \mathcal{W} to be the minimum hop-distance between node v and any node in \mathcal{W} , i.e.,

$$d(v, \mathcal{W}) \triangleq \min_{u \in \mathcal{W}} d(v, u).$$

We then define the *infection eccentricity* of node set \mathcal{W} to be the maximum hop-distance from an infected node in \mathcal{I}' to set \mathcal{W} , i.e.,

$$e(\mathcal{W}, \mathcal{I}') = \max_{v \in \mathcal{I}'} d(v, \mathcal{W}). \quad (1)$$

We further define m -Jordan-cover (m -JC) to be the set $\mathcal{W}^*(\mathcal{K}, \mathcal{I}', m)$ such that

$$\mathcal{W}^*(\mathcal{K}, \mathcal{I}', m) = \operatorname{argmin}_{\mathcal{W} \in \{\mathcal{W} | |\mathcal{W}|=m, \mathcal{W} \subset \mathcal{K}\}} e(\mathcal{W}, \mathcal{I}'). \quad (2)$$

Algorithm 1: The Candidate Selection Algorithm

Input: \mathcal{I}', g, Y ;

Output: g^- (the candidate subgraph)

Set \mathcal{K} to be an empty set.

for $v \in \mathcal{V}(g)$ **do**

if $|\mathcal{N}(v) \cap \mathcal{I}'| \geq Y$ **then**

 Add v to \mathcal{K} , where $\mathcal{N}(v)$ is the set of neighbors of node v .

end

end

Set \mathcal{K}^+ to be $\mathcal{K} \cup \mathcal{I}'$.

Set g' to be the graph induced by set \mathcal{K}^+ .

Find all connected components in g' .

if g' is connected **then**

 Set $g^- = g'$.

else

 Randomly select one node in each components of g' . Denote by \mathcal{R} the set of the selected nodes.

 Randomly select one node $v \in \mathcal{R}$.

for $u \in \mathcal{R} \setminus v$ **do**

 Compute the shortest path P from v to u .

 Set $g' = g' \cup P$.

end

 Set $g^- = g'$

end

return g^-, \mathcal{K} .

where \mathcal{I}' is the set of observed infected nodes that m -JC needs to cover and \mathcal{K} is the *candidate set* for the sources. Therefore, m -JC is the set of m nodes in \mathcal{K} with the minimum infection eccentricity.

We now introduce the *optimal Jordan cover* (OJC) algorithm whose asymptotic detection rate will be analyzed in Section 4.1.

The Optimal Jordan Cover (OJC) Algorithm

- **Step 1: Candidate Selection:** Let Y be a positive integer. The candidate set \mathcal{K} is the set of nodes with more than Y observed infected neighbors. In addition, define $\mathcal{K}^+ \triangleq \mathcal{K} \cup \mathcal{I}'$. Denote by g^- a connected subgraph of g induced by node set \mathcal{K}^+ . An induced graph is a subset of nodes of a graph with all edges whose endpoints are both in the node subset. If the induced graph is not connected, we select a random node in each component, randomly pick one selected node and add the shortest paths from this node to all other selected nodes to form a connected g^- . We call Y the *selection threshold*. The pseudo code of the candidate selection algorithm for selecting \mathcal{K} and g^- can be found in Algorithm 1.
- **Step 2: Jordan Cover:** For any m combination of nodes in \mathcal{K} in Step 1, we compute the infection eccentricity of the node set as defined in (1) on subgraph g^- , and select the combination with the minimum infection eccentricity as the set of sources. Ties are broken by the total distance from the observed infected to the node set, i.e., $\sum_{v \in \mathcal{I}'} d(v, \mathcal{W})$.

With a properly chosen threshold Y , the candidate selection step includes all sources in \mathcal{K} with a high probability and excludes nodes that are more than $t + 1$ hops away from all sources. By limiting the computation on the induced subgraph g^- , the computational complexity is reduced significantly. From simulations, we will see that it results in $27\times$ reduction of the running time without affecting the detection rate. The asymptotic detection rate of OJC will be studied in Theorem 2. Under some conditions, OJC identifies all sources with probability one asymptotically.

OJC is a polynomial-time algorithm for given m , but the complexity increases exponentially in m . To further reduce the complexity, we propose Approximate Jordan Cover (AJC), which replaces Step 2 of OJC with the K-Means algorithm [7]. As shown in the simulations, the performance of the AJC algorithm, in terms of both detection rate and the error distance, is close to OJC with much shorter running time. The computational complexity of both algorithms are summarized in the following theorem.

Theorem 1. *The computational complexity of OJC is*

$$O \left(|\mathcal{I}'| \left(|\mathcal{E}(g)| + m \binom{|\mathcal{V}(g^-)|}{m} \right) \right),$$

and the computational complexity of AJC is

$$O \left(|\mathcal{I}'| (|\mathcal{E}(g)| + H|\mathcal{V}(g^-)|) \right),$$

where H is the number of iterations used in the K-Means algorithm in AJC.

Proof. We first analyze the complexity of the OJC algorithm. For simplicity, denote by $V = |\mathcal{V}(g)|$ the number of nodes and $E = |\mathcal{E}(g)|$ the number of edges.

In the candidate selection stage, each node needs to compute its degree. Therefore, each edge is counted twice and each node is processed once. The complexity is $O(V + E)$. Counting the number of the connected components using breadth first search is of complexity $O(V + E)$. When the induced graph is not connected,

the complexity to compute the shortest paths from one node to all other nodes in an unweighted graph is of complexity $O(V + E)$. As a summary, the complexity of the candidate selection algorithm is $O(V + E)$.

The resulting subgraph has $|\mathcal{V}(g^-)|$ nodes and at most E edges. Next, we compute the complexity of the OJC.

We first compute the distances from nodes in g^- to nodes in \mathcal{I}' . Note this is equivalent to do a breadth-first search from each node in \mathcal{I}' . The complexity of the BFS is $O(|\mathcal{V}(g^-)| + E)$. Therefore, the complexity for this step is $O(|\mathcal{I}'|(|\mathcal{V}(g^-)| + E))$. The results are saved in a two dimensions hashtable so that querying the distance from one observed infected node and another node in graph g^- is of complexity $O(1)$.

After the above precomputation, for each set of nodes with size m , we want to obtain its infection eccentricity. For each observed infected node, we query the hash table to find the minimum distance to the set of nodes with size m . The complexity is $O(m)$. To compute the infection eccentricity of one set is of complexity $O(m|\mathcal{I}'|)$. In addition, there are $\binom{|\mathcal{V}(g^-)|}{m}$ possible node sets. Therefore, the complexity is $O((m|\mathcal{I}'|)\binom{|\mathcal{V}(g^-)|}{m})$.

As a summary, the complexity of the OJC algorithm is

$$\begin{aligned} & O \left(V + E + |\mathcal{I}'|(|\mathcal{V}(g^-)| + E) + m|\mathcal{I}'| \binom{|\mathcal{V}(g^-)|}{m} \right) \\ & = O \left(|\mathcal{I}'| \left(E + m \binom{|\mathcal{V}(g^-)|}{m} \right) \right) \end{aligned}$$

Next, we analyze the complexity of the AJC algorithm. Note the complexity of the candidate selection and the precomputation are the same. The complexity of the Kmeans algorithm is analyzed as follows.

The complexity of the membership assignment phase is $O(m|\mathcal{I}'|)$. For each observed infected node, we only need to query its distance to the preselected m sources and each query is of complexity $O(1)$ based on the results of the precomputation.

For the center update phase, for each cluster, we need to search all $|\mathcal{V}(g^-)|$ nodes to find a center. Therefore, the complexity is $|\mathcal{V}(g^-)||\mathcal{I}'|$.

As a summary, the complexity of one iteration of the Kmeans algorithm is

$$O((m + |\mathcal{V}(g^-)|)|\mathcal{I}'|)$$

Denote by H the number of iterations, the complexity of the AJC algorithm is

$$\begin{aligned} & O(V + E + |\mathcal{I}'|(|\mathcal{V}(g^-)| + E) + N((m + |\mathcal{V}(g^-)|)|\mathcal{I}'|)) \\ & = O(|\mathcal{I}'|(E + H|\mathcal{V}(g^-)|)) \end{aligned}$$

□

4 Asymptotic Analysis of OJC

In this section, we present the asymptotic analysis of the detection rate of OJC on the ER random graph. The results include the conditions that guarantee probability one detection and the conditions under which it is impossible to detect the source set with nonzero probability under any source localization algorithm.

4.1 Asymptotic perfect detection on the ER random graph

We first present the positive result that shows that on the ER random graph, OJC identifies the m sources with probability one asymptotically under some conditions. Recall that n is the number of nodes in the graph, and m is the number of sources, which is a constant independent of n . Denote by p the wiring probability of the ER random graph, which is the probability that there exists a link between two nodes. Let $\mu = np$, which is the average node degree. Define $q \triangleq \min_{u,v \in \mathcal{V}(g)} q_{u,v}$, i.e., the minimum infection probability over all edges and $\theta \triangleq \min_{v \in \mathcal{V}(g)} \theta_v$ i.e., the minimum report probability over all nodes.

Theorem 2. *OJC identifies all m sources with probability one as $n \rightarrow \infty$ when the following conditions hold:*

- (c1): $\mu q \theta = \Omega(\log n)$,¹
- (c2): $\limsup_{n \rightarrow \infty} \frac{Y}{\mu q \theta} < 1$ and $\liminf_{n \rightarrow \infty} \frac{Y}{\mu q \theta} > 0$, and
- (c3): $t = \omega(D)$ and $\limsup_{n \rightarrow \infty} \frac{t}{\frac{\log n}{\log \mu}} < \frac{2}{3}$.

Proof. In this proof, we will show that one source $s_i \in \mathcal{W}^*(\mathcal{K}, \mathcal{I}', m)$ with a high probability. Then with a union bound, we will show that

$$\mathcal{S} = \mathcal{W}^*(\mathcal{K}, \mathcal{I}', m).$$

with a high probability for sufficiently large n .

Recall that we regard the recovered nodes and infected nodes as "infected" since the recovery process is not related to the proof. Without loss of generality, we consider s_1 . Throughout the proof, we consider the BFS tree rooted at s_1 . In particular, the level of one node means the level of the node on the BFS tree rooted at s_1 .

We first introduce and recall some necessary notations terms.

For an ER random graph g .

- A node v is said to be on level i if $d_{s_1 v} = i$. Denote by \mathcal{L}_i the set of nodes from level 0 to level i and $l_i = |\mathcal{L}_i|$.
- Denote by L'_i the set of nodes on level i . In addition, l'_i is the number of nodes on level i .
- The descendants of node v in a tree are all the nodes in the subtree rooted at v . In addition, v is the ancestor of all its descendants.
- The offsprings of a node on level k (say v) are the nodes which are on level $k + 1$ and have edges to v . Denote by $\Phi(v)$ the offspring set of v and $\phi(v) = |\Phi(v)|$.
- Denote by p the wiring probability in the ER random graph.
- Denote by n the total number of nodes.
- Denote by $\mu = np$.
- Denote by $\text{Bi}(n, p)$ the binomial distribution with n number of trials and each trial succeeds with probability p .

¹Throughout this paper, the asymptotic order notation is defined for $n \rightarrow \infty$.

- Denote by T^\dagger the BFS tree rooted at s_1 .
- Denote by $\Phi'(v)$ the set of offsprings of node v on T^\dagger and $\phi'(v) = |\Phi'(v)|$.
- Denote by g_t the subgraph induced by all nodes within t hops from s on the ER graph. The *collision edges* are the edges which are not in T^\dagger but in g_t , i.e., $e \in \mathcal{E}(g_t) \setminus \mathcal{E}(T^\dagger)$.
- A node who is an end node of a collision edge is called a *collision node*. Denote by \mathcal{R}_k the set of collision edges whose end nodes are within level k and $R_k = |\mathcal{R}_k|$.
- Denote by \mathcal{Z}_i the set of nodes which are infected at time i .
- Denote by $\tilde{\mathcal{Z}}_j^i(v)$ the set of nodes that are infected at time slot i , on level j , when s_1 is the only infection source in the graph and the descendants of node v in the BFS tree rooted at s_1 and $\tilde{\mathcal{Z}}_j^i(v)$ are the observed infected nodes in $\tilde{\mathcal{Z}}_j^i(v)$. $\tilde{Z}_j^i(v)$ and $\tilde{Z}_j^i(v)$ are defined as the cardinality respectively.
- Denote by $\psi(v)$ the number of observed infected neighbors of node v .
- Denote by $\psi'(v)$ the number of infected offsprings of node v (the offspring is defined based on the BFS tree rooted at node s_1).
- Denote by $\psi''(v)$ the number of *observed* infected offsprings of node v .

To prove $s_1 \in \mathcal{W}^*(\mathcal{K}, \mathcal{I}', m)$, we need to show that any set \mathcal{W} such that $s_1 \notin \mathcal{W}$ has a infection eccentricity larger than t on g^- . We need the following asymptotic high probability events.

- **Offsprings of each node.** Consider the BFS tree rooted at source s_1 . Define

$$E_1 = \{\forall v \in \mathcal{L}_{t+D}, \phi'(v) \in ((1-\delta)\mu, (1+\delta)\mu)\}.$$

E_1 , when occurs, provides upper and lower bounds for the number of offsprings of each node in \mathcal{L}_{t+D} .

- **Total number collision edges.** Consider the BFS tree rooted at source s_1 . We define event E_2 when the following upper bound on the collision edges holds

$$R_j \begin{cases} = 0 & \text{if } 0 < j \leq \lfloor m^- \rfloor, \\ \leq 8\mu & \text{if } \lfloor m^- \rfloor < j < \lceil m^+ \rceil, \\ \leq \frac{4[(1+\delta)\mu]^{2j+1}}{n} & \text{if } \lceil m^+ \rceil \leq j \leq \frac{\log n}{(1+\alpha)\log \mu}. \end{cases}$$

where $m^+ = \frac{\log n}{2 \log[(1+\delta)\mu]}$ and $m^- = \frac{\log n - 2 \log \mu - \log 8}{2 \log[(1+\delta)\mu]}$. E_2 provides the upper bounds for collision edges at different levels. Note that a subgraph with diameter $\leq m^-$ is a tree with high probability since there is no collision edges.

- **Detailed collision edges in level $> D + t$.** Define E_3 to be the event that

$$\forall v \in \mathcal{L}'_{D+t+1}, \quad \psi(v) < (1-\delta)^3 \mu q \theta.$$

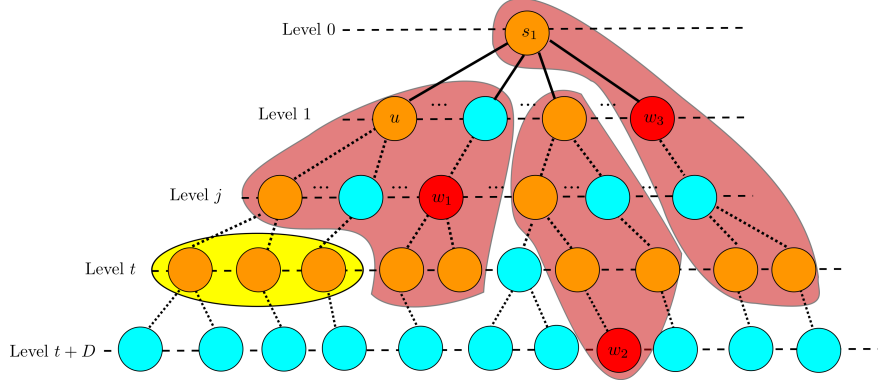


Figure 1: A pictorial example for Theorem 2.

- **Infected nodes.** Define

$$E_4 = \{\forall v \in \cup_{i=0}^{t-1} \mathcal{Z}_i, \quad \psi'(v) > (1 - \delta)^2 \mu q.\}$$

and

$$E_5 = \{\forall v \in \cup_{i=0}^{t-1} \mathcal{Z}_i, \quad \psi''(v) > (1 - \delta)^3 \mu q \theta.\}$$

- **Infected nodes from source s_1 .** Define

$$E_6 = \{\tilde{Z}_1^1 \geq (1 - \delta)^2 \mu q\} \cap \{\forall v \in \tilde{Z}_1^1, \tilde{Z}_t^t(v) \geq [(1 - \delta)^2 \mu q]^{t-1} (1 - \delta) \theta\}$$

To prove these event happens with a high probability, we have

$$\begin{aligned} & \Pr(E_1 \cap E_2 \cap E_3 \cap E_4 \cap E_5 \cap E_6) \\ & \geq \Pr(E_1)(1 - \Pr(\bar{E}_2|E_1) - \Pr(\bar{E}_3|E_1) - \Pr(\bar{E}_6|E_1)) - \Pr(\bar{E}_4 \cup \bar{E}_5) \\ & = \Pr(E_1)(1 - \Pr(\bar{E}_2|E_1) - \Pr(\bar{E}_3|E_1) - \Pr(\bar{E}_6|E_1)) - (1 - \Pr(E_4 \cap E_5)) \\ & \geq 1 - \epsilon, \end{aligned}$$

for sufficiently large n . Based on Lemma 1, 2, 4 and 5, we have events $E_1, E_2, E_3, E_4, E_5, E_6$ happen with a high probability as n is large enough. The proofs can be found in the Appendices.

We summarize the important properties of the OJC algorithm based on the above asymptotic events.

- Let $Y = (1 - \delta)^3 \mu q \theta$. In Step 1 of the OJC algorithm, we have $\mathcal{K} \subset \mathcal{L}_{t+D}$, i.e., all nodes on the subgraph g^- are within level $t+D$ of the BFS tree rooted at source s_1 . Based on E_3 , all nodes on level $t+D+1$ have less than Y observed infected neighbors. In addition, all nodes one level $l > t+D+1$ have no infected neighbors since all the infected nodes are within level $t+D$.
- Based on event E_5 , all nodes which are infected at or before time $t-1$ have at least Y observed infected neighbors. Therefore, we have $\cup_{i=0}^{t-1} \mathcal{Z}_i \subset \mathcal{K}$ which implies $\mathcal{I} \subset \mathcal{K}^+$ and g^- is a connected graph.

Therefore, g^- is a connected graph which contains all the infected nodes and is restricted up to level $t+D$ based on E_3 and E_5 .

Next we will show that $s_1 \in \mathcal{W}^*$ by contradiction. Note, we have $e(\mathcal{S}, \mathcal{I}') = t$. Therefore, $e(\mathcal{W}^*, \mathcal{I}') \leq t$. We will show that $e(\mathcal{W}, \mathcal{I}') > t$ for any set of nodes \mathcal{W} where $|\mathcal{W}| = m$, $s_1 \notin \mathcal{W}$. Note, all the infection eccentricity are considered based on the subgraph g^- . Specifically, we will show that no nodes in g^- other than source s_1 can reach all nodes which are infected by source s_1 within t time slot based on events E_1, E_2, E_3, E_4, E_5 and E_6 .

Consider the BFS tree rooted at source s_1 . We discuss the proof in two cases.

- (a) When $t + D \leq \lfloor m^- \rfloor$, according to the event E_2 , the $t + D$ hops from s_1 is a tree because there is not collision edges. When event E_6 occur, there are at least $(1 - \delta)^2 \mu q$ infected nodes at level 1 and $\forall v \in \mathcal{Z}_1^1, Z_t^{tt}(v) \geq [(1 - \delta)^2 \mu q]^{t-1} (1 - \delta) \theta$ which means there exists at least one observed infected node on level t for each subtree rooted at level 1. Consider a set \mathcal{W} where $s_1 \notin \mathcal{W}$. There exists a node $u \in \tilde{\mathcal{Z}}_1^1$ such that $\mathcal{W} \cap \mathcal{V}(T_u^{-s_1}) \neq \emptyset$ where $T_u^{-s_1}$ is the tree rooted at node u without the branch of node s_1 (the subtree on level 1). Therefore, for one node $w \in \tilde{\mathcal{Z}}_t^{tt}(u)$, we have $d(w, \mathcal{W}) > t$. Hence for any set \mathcal{W} with size m that does not contain s_1 , $e(\mathcal{W}, \mathcal{I}') > t$. Therefore, we have $s_1 \in \mathcal{W}^*$ when $t + D \leq \lfloor m^- \rfloor$.

Figure 1 shows a pictorial example when $m = 3$. The red nodes are the nodes in set \mathcal{W} . The existence of node u is guaranteed since m nodes are insufficient to cover all level 1 branches of the BFS tree rooted in s_1 . The red areas are the t hop neighborhood of nodes w_1, w_2 and w_3 . In this case, the $t + D$ neighborhood of node s_1 is a tree. Therefore, any set \mathcal{W} does not contain s_1 can not reach the yellow area $\tilde{\mathcal{Z}}_t^{tt}(u)$ within t hops as shown in the figure.

- (b) Consider the case when $t + D > \lfloor m^- \rfloor$. Again, based on event E_6 , there exists a node $u \in \tilde{\mathcal{Z}}_1^1$ such that $\mathcal{W} \cap \mathcal{V}(T_u^{-s_1}) \neq \emptyset$. The distance between a node in $\mathcal{Z}_t^{tt}(u)$ and set \mathcal{W} on the BFS tree is larger than t . Therefore, if \mathcal{W} is a Jordan infection cover, the shortest paths between $\mathcal{Z}_t^{tt}(u)$ and set \mathcal{W} must contain at least one collision nodes. In the rest of the proof, we will show that the number of collision nodes is insufficient to provides the “shortcuts” to all observed infected nodes.

Define H to be the total number of nodes each of which has the shortest path to \mathcal{W} within t hops and containing at least one collision node. If $H < \tilde{\mathcal{Z}}_t^{tt}(u)$, there exists a node $w \in \mathcal{Z}_t^{tt}(u)$ such that $d(w, \mathcal{W}) > t$. Therefore, \mathcal{W} can not be the Jordan infection cover and the theorem is proved.

In the rest of the proof, we will show that $H < \tilde{\mathcal{Z}}_t^{tt}(u)$. We first have the lower bound on $\tilde{\mathcal{Z}}_t^{tt}(u)$ according to E_6 ,

$$\tilde{\mathcal{Z}}_t^{tt}(u) \geq [(1 - \delta)^2 \mu q]^{t-1} (1 - \delta) \theta \quad (3)$$

For each node w_i in \mathcal{W} , define H_i to be the total number of nodes each of which has the shortest path to w_i within t hops and containing at least one collision node. The upper bound of H_i can be obtained based on Lemma 6 in [23].

$$H_i \leq c[(1 + \delta)\mu]^{\frac{3}{4}(t+D)+\frac{1}{2}} + c[(1 + \delta)\mu]^{(\frac{5}{4}-\frac{\alpha}{2})(t+D)+2},$$

and we have

$$H \leq \sum_{i=2}^m H_i \leq mc[(1 + \delta)\mu]^{\frac{3}{4}(t+D)+\frac{1}{2}} + mc[(1 + \delta)\mu]^{(\frac{5}{4}-\frac{\alpha}{2})(t+D)+2},$$

Since $\frac{1}{2} < \alpha < 1$, we have $\alpha = \frac{1}{2} + \alpha'$ where $0 < \alpha' < \frac{1}{2}$ is a constant, we have

$$\frac{H}{\tilde{\mathcal{Z}}_t^{tt}(u)} \leq \frac{mc[(1 + \delta)\mu]^{\frac{3}{4}(t+D)+\frac{1}{2}}}{[(1 - \delta)^2 \mu q]^{t-1} (1 - \delta) \theta} + \frac{mc[(1 + \delta)\mu]^{(\frac{5}{4}-\frac{\alpha}{2})(t+D)+2}}{[(1 - \delta)^2 \mu q]^{t-1} (1 - \delta) \theta} \quad (4)$$

$$= \frac{mc}{\mu} \left(\frac{(1+\delta)^{\frac{3}{4} + (\frac{3D}{4} + \frac{1}{2})\frac{1}{t}}}{[(1-\delta)^2 q]^{1-\frac{1}{t}} (1-\delta) \theta \mu^{\frac{1}{4} - (\frac{3D}{4} + \frac{5}{2})\frac{1}{t}}} \right)^t \quad (5)$$

$$+ \frac{mc}{\mu} \left(\frac{(1+\delta)^{1-\frac{\alpha'}{2} + [(1-\frac{\alpha'}{2})^{D+2}]^{\frac{1}{t}}}}{[(1-\delta)^2 q]^{1-\frac{1}{t}} (1-\delta) \theta \mu^{\frac{\alpha'}{2} - (4 + (1-\frac{\alpha'}{2})D)\frac{1}{t}}} \right)^t \quad (6)$$

Since $t \gg D$, we have

$$t \geq \min \left\{ 3D + 2, \left(\frac{2}{\alpha'} - 1 \right) D + \frac{4}{\alpha'}, \left(\frac{4}{\alpha'} - 2 \right) D + \frac{16}{\alpha'} \right\}$$

Therefore, Inequality (6) becomes

$$\frac{H}{\tilde{Z}_t^t(u)} \leq \frac{2mc}{\mu} \left(\frac{(1+\delta)}{[(1-\delta)^2 q] \mu^{\frac{\alpha'}{4}}} \right)^t.$$

Since $\mu > \frac{1}{Cq\theta} \log n$ and δ, q, α' are constants, we have

$$\frac{(1+\delta)}{[(1-\delta)^2 q] \mu^{\frac{\alpha'}{4}}} < 1$$

when

$$n > \exp \left(\frac{1}{2} \left(\frac{(1+\delta)}{(1-\delta)^2 q} \right)^{\frac{4}{\alpha'}} \right).$$

Therefore, we have

$$\frac{H}{\tilde{Z}_t^t(u)} \leq \frac{2mc}{\mu} \leq \epsilon',$$

where $\epsilon' \in (0, 1)$ is a constant and the inequality holds for sufficiently large n . There are at least $(1 - \epsilon') \tilde{Z}_t^t(u)$ nodes which cannot be reached from \mathcal{W} with t hops on g^- . Hence we have $e(\mathcal{I}', \mathcal{W}) > t$. Therefore, we proved that $s_1 \in \mathcal{W}^*$ with a high probability when n is sufficiently large, i.e., we have

$$\Pr(s_1 \in \mathcal{W}^*) \geq 1 - \frac{\epsilon}{m}$$

since m is a constant. Then, by applying the union bound, we have, for sufficiently large n ,

$$\Pr(s_i \in \mathcal{W}^*, \quad \forall i = 1, \dots, m) \geq 1 - \epsilon$$

Note, we have $|\mathcal{W}^*| = m$. Therefore, we have

$$\Pr(\mathcal{S} = \mathcal{W}^*) \geq 1 - \epsilon$$

Hence, the Jordan infection cover equals the actual source set with a high probability. \square

We now briefly explain the conditions. Recall that μ is the average node degree, q is the lower bound on the infection probability and θ is the lower bound on the reporting probability, so $\mu q \theta$ is a lower bound on the average number of observed infected neighbors of a node that was infected before time slot t . Therefore, condition (c1) requires that this lower bound is $\Omega(\log n)$, and condition (c2) requires that the threshold used in the candidate selection algorithm is a constant fraction of the average number of observed infected neighbors. Applying the Chernoff bound, conditions (c1) and (c2) together yield the following conclusions:

- (i) any node who was infected before or at time slot $t - 1$ (hence, including the sources) will be selected into the candidate set with a high probability,
- (ii) any node that is $t + D + 1$ hops away from the set of sources will not have Y or more observed infected neighbors with a high probability, and
- (iii) any node that is more than $t + D + 1$ hops away from the set of sources will not have any observed infected neighbors.

Based on the above facts, with a high probability, the candidate set includes all nodes who were infected at $t - 1$ or earlier, and any node in g^- is at most $t + D$ hops away from all sources.

Condition (c3) is on the infection duration. We first restrict $t = \omega(D)$ so that the infection subgraphs starting from different sources are likely to overlap and form a connected component. This is a more interesting regime than the one in which infection subgraphs are disconnected from each other. $\limsup_{n \rightarrow \infty} \frac{t}{\frac{\log n}{\log \mu}} < \frac{2}{3}$ is a critical condition. The intuition why it is required is explained below. Figure 1 provides a pictorial explanation of the proof. The picture illustrates the breadth-first-search (BFS) tree T^\dagger rooted at source s_1 with height $t + D$, where s_1 is one of the m sources. The nodes in orange are the observed infected nodes whose infection was originated from s_1 . The blue nodes are unobserved nodes. A node is said to be on level i of the BFS if its hop-distance to s_1 is i . Assume $m = 3$ and consider a set of three nodes who are within $t + D$ hops from s_1 but not includes s_1 (e.g., $\mathcal{W} = \{w_1, w_2, w_3\}$ in Figure 1). Suppose the infection eccentricity of \mathcal{W} is $\leq t$. Since s_1 has a sufficient number of neighbors according to the definition of μ , with a high probability, there exists a subtree of T^\dagger starting from an offspring of s_1 , which does not include any node in \mathcal{W} . Assume u is the root of such a subtree in Figure 1. The yellow area in Figure 1 includes the level- t observed infected nodes on subtree $T_u^{-s_1}$. Any path from w_1, w_2 or w_3 to the yellow area, formed by edges on T^\dagger , must have hop-distance larger than t . Therefore, if the infection eccentricity of \mathcal{W} is at least t , there must exist a path from \mathcal{W} to each of the nodes in the yellow area with hop-distance $\leq t$, and such a path must include at least one edge which is not in T^\dagger (we call these edges *collision edges*). In the detailed analysis, we will prove that with a high probability, the number of nodes within t hops from \mathcal{W} via the collision edges is order-wise smaller than the number of nodes in the yellow area when (c3) holds. Therefore, the Jordan cover has to include s_1 . The same argument applies to other sources.

4.2 Impossibility results

Theorem 5 in [24] presents the conditions under which it is impossible to identify the single source under the IC diffusion on the ER random graph, which is a special case of the model in this paper. Assuming SI or IC model, based on Lemma 1 in [24], we have that with a high probability, all nodes of the ER graph become infected when the infection duration is at least

$$t_u \triangleq \left\lceil \frac{\log n}{\log \mu + \log q} \right\rceil + 2.$$

When this occurs, it is impossible to detect the sources since the nodes are indistinguishable.

Theorem 3. *Assume the multi-source diffusion follows the IC or SI model. If $24 \log n < q\mu < \sqrt{n}$ and q is a constant independent of n , then*

$$\lim_{n \rightarrow \infty} \Pr(\mathcal{I} = \mathcal{V}(g)) = 1$$

when the observation time $t \geq t_u$. In other words, the entire network is infected after t_u with a high probability. In such a case, the probability of finding the sources diminishes to zero as $n \rightarrow \infty$.

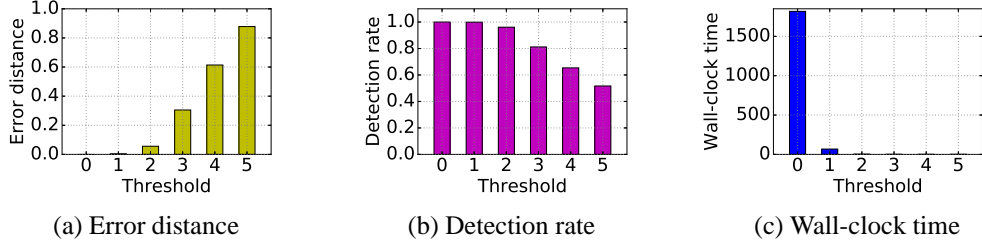


Figure 2: Performance of OJC with different threshold values on the ER random graph

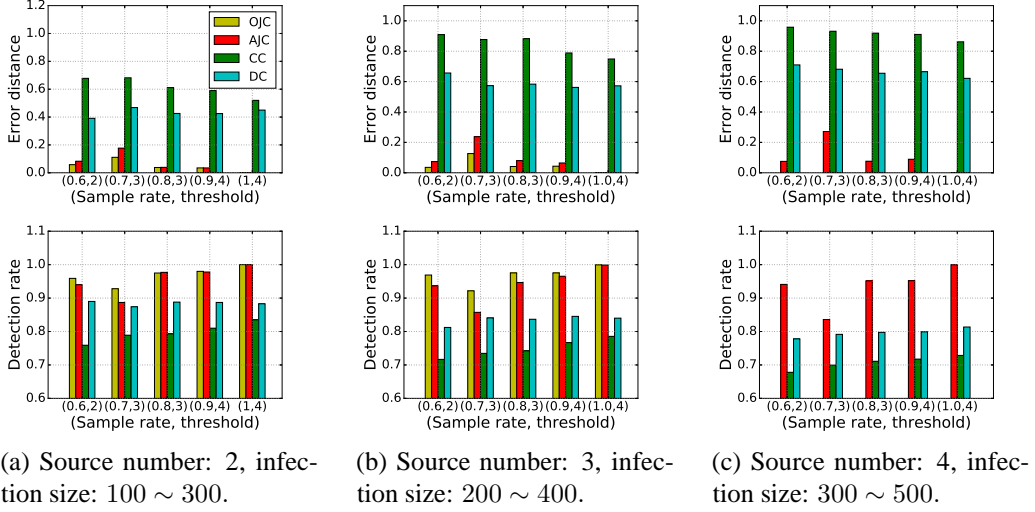


Figure 3: The Performance of OJC, AJC, CC and DC on the ER random graph with different sample rates and threshold values

5 Simulations

In this section, we evaluated the performance of our algorithms via simulations. The performance metrics used in this paper include:

- Error distance: The error distance is defined to be

$$\min_{\mathcal{P} \in \text{permutation}(\mathcal{S}')} \sum_{i=1}^m \frac{d(s_i, p_i)}{m},$$

where s_1, s_2, \dots, s_m are the real sources, \mathcal{S}' is the set of detected sources and $\mathcal{P} = (p_1, p_2, \dots, p_m)$ is a permutation of \mathcal{S}' .

- Detection rate: The detection rate is defined as

$$\frac{|\mathcal{S} \cap \mathcal{S}'|}{m}.$$

We compared our algorithms with two heuristic algorithms (DC and CC) based on K-Means, which have been used for comparison in [12]. The algorithms proposed in [12] and [3] are the same as AJC without

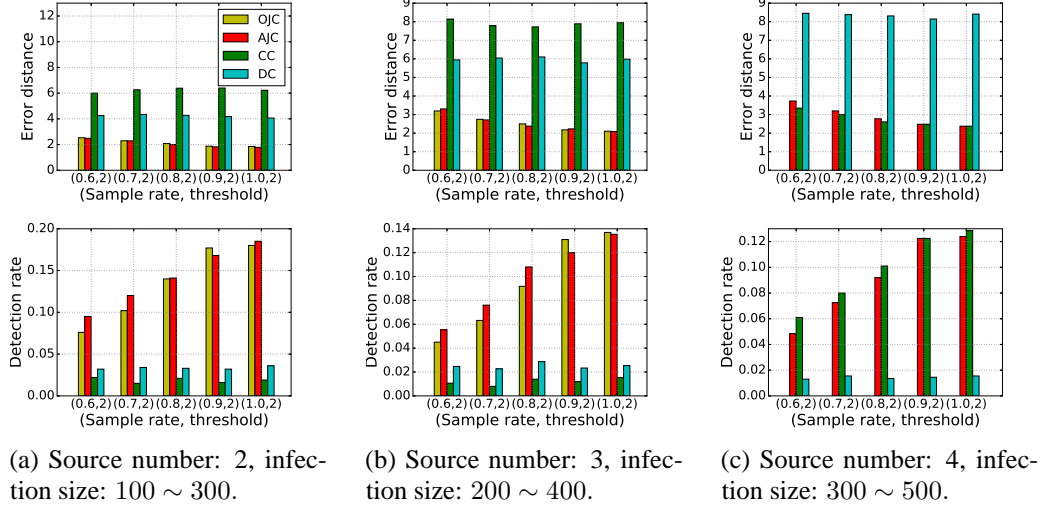


Figure 4: The Performance of OJC, AJC, CC and DC on the power grid network with different sample rates and threshold values

candidate selection. In both DC and CC, the initial centroids are randomly chosen. During the clustering step of each iteration in K-Means, we selected distance centroid of each cluster in DC and selected closeness centroid in CC, where distance centroid is defined as $\arg \min_{v \in \mathcal{C}} \sum_{u \in \mathcal{C} \cap \mathcal{I}'} d(v, u)$, where \mathcal{C} is the set of nodes in the cluster. Closeness centroid is defined as $\arg \max_{v \in \mathcal{C}} \sum_{u \in \mathcal{C} \cap \mathcal{I}', u \neq v} \frac{1}{d(u, v)}$. The following experiments were conducted on a server with 8 Intel Xeon X3450 CPUs and 16G RAM with Linux 64 bit system. All algorithms were implemented with Python 2.7.

5.1 OJC with different thresholds

In Figure 2, we evaluated OJC on the ER random graph. In the experiments, we generated an ER random graph with 5,000 nodes and wiring probability 0.002. We used the homogeneous SI model for diffusion with infection probability 0.8. In this experiment, we limited the infection network size to be 100 ~ 300 and the number of sources to be 2 due to the computational complexity of the OJC algorithm. Figure 2 shows the performance of OJC with different thresholds. From the results, the detection rate is close to one and the error distance is close to zero under OJC with threshold 0 or 1. However, the running time is 1,817 seconds versus 68 seconds. So the candidate selection algorithm with threshold one results in $27\times$ reduction of the running time. When the threshold increases 2, the running time reduces to 3 seconds, which is a $600\times$ reduction of the running time. Both the detection rate and the error distance became slightly worse in this case. The detection rate in this case is 0.961 and the error distance is 0.056.

5.2 OJC, AJC and other heuristics

We further evaluated the performance of OJC and AJC on both the power grid network [19] and ER random graph (size: 5000, wiring probability: 0.002) and compared them with DC and CC heuristics. We used the homogeneous SI model with infection probability 0.8 to generate the diffusion sequences. For AJC/CC/DC, for each diffusion sequence, we repeated the algorithm 100 times from different initial conditions and chose the source set with the smallest the smallest-infection-eccentricity/largest-closeness-

centrality/smallest-distance-centrality. In Figure 3 and 4, the x -axis represents the combinations of sample rate and threshold. On the ER random graph, we increased the threshold as the sample rate increased to control the running time. For the power-grid network, since the average node degree is only 2, we set threshold equal to 2 for experiments for all sample rates. As we can see from the figures that when fixing the threshold, the performance of all algorithms (in terms of both error distance and detection rate) improves as the sample rate increases because we had more information about the diffusion. From Figure 3 and 4, we can also see that AJC outperforms DC and CC, and has similar performance with OJC. Note that with four sources, OJC became very slow on both the ER random graph and the power grid network because its complexity increases exponentially in the number of sources. So for the cases with four sources, we only simulated AJC.

6 Conclusions

In this paper, we studied the problem of detecting multiple diffusion sources under the heterogeneous SIR model with incomplete observations. We defined a concept called Jordan cover and developed the OJC algorithm based on that. Our theoretical analysis showed that OJC finds the set of sources in the ER random graph with probability one asymptotically under mild conditions. To the best of our knowledge, this is the first theoretic performance guarantee for multiple information sources detection in non-tree networks. Since the computational complexity of OJC is polynomial in n but exponential in m , we proposed a heuristic algorithm — the AJC algorithm. Our simulation results showed that OJC and AJC algorithms have similar performance and both significantly outperform existing algorithms.

7 Acknowledgments

This work was supported in part by the U.S. Army Research Laboratory’s Army Research Office (ARO Grant No. W911NF1310279) and by the Defense Threat Reduction Agency (DTRA Grant HDTRA1-16-0017).

References

- [1] Ameya Agaskar and Yue M Lu. A fast monte carlo algorithm for source localization on graphs. In *SPIE Optical Engineering and Applications*, 2013.
- [2] N. T. J. Bailey. *The mathematical theory of infectious diseases and its applications*. Hafner Press, 1975.
- [3] Zhen Chen, Kai Zhu, and Lei Ying. Detecting multiple information sources in networks under the SIR model. In *Proc. IEEE Conf. Information Sciences and Systems (CISS)*, Princeton, NJ, 2014.
- [4] P. Erdos and A. Renyi. On random graphs I. *Publ. Math. Debrecen*, 6:290–297, 1959.
- [5] Mehrdad Farajtabar, Manuel Gomez-Rodriguez, Mohammad Zamani, Nan Du, Hongyuan Zha, and Le Song. Back to the past: Source identification in diffusion networks from partially observed cascades. In *Int. Conf. Artificial Intelligence and Statistics*, San Diego, CA, May 2015.

- [6] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 12(3):211–223, 2001.
- [7] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.
- [8] Nikhil Karamchandani and Massimo Franceschetti. Rumor source detection under probabilistic sampling. In *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Istanbul, Turkey, July 2013.
- [9] T. Lappas, E. Terzi, D. Gunopulos, and H. Mannila. Finding effectors in social networks. In *Proc. Ann. ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD)*, pages 1059–1068, 2010.
- [10] Andrey Y. Lokhov, Marc Mézard, Hiroki Ohta, and Lenka Zdeborová. Inferring the origin of an epidemic with a dynamic message-passing algorithm. *Phys. Rev. E*, 90, Jul 2014.
- [11] Wuqiong Luo, Wee Peng Tay, and Mei Leng. Identifying infection sources and regions in large networks. *IEEE Trans. Signal Process.*, 61:2850–2865, 2013.
- [12] Wuqiong Luo, Wee Peng Tay, and Mei Leng. On the universality of jordan centers for estimating infection sources in tree networks. *arXiv preprint arXiv:1411.2370*, 2014.
- [13] Pedro C Pinto, Patrick Thiran, and Martin Vetterli. Locating the source of diffusion in large-scale networks. *Phys. Rev. Lett.*, 109(6):068702, 2012.
- [14] B. A. Prakash, J. Vreeken, and C. Faloutsos. Spotting culprits in epidemics: How many and which ones? In *IEEE Int. Conf. Data Mining (ICDM)*, pages 11–20, Brussels, Belgium, 2012.
- [15] D. Shah and T. Zaman. Detecting sources of computer viruses in networks: Theory and experiment. In *Proc. Ann. ACM SIGMETRICS Conf.*, pages 203–214, New York, NY, 2010.
- [16] D. Shah and T. Zaman. Rumors in a network: Who’s the culprit? *IEEE Trans. Inf. Theory*, 57:5163–5181, Aug. 2011.
- [17] D. Shah and T. Zaman. Rumor centrality: a universal source detector. In *Proc. Ann. ACM SIGMETRICS Conf.*, pages 199–210, London, England, UK, 2012.
- [18] Z. Wang, W. Dong, W. Zhang, and C. W. Tan. Rumor source detection with multiple observations: fundamental limits and algorithms. In *Proc. Ann. ACM SIGMETRICS Conf.*, Austin, TX, 2014.
- [19] Duncan J Watts and Steven H Strogatz. Collective dynamics of ‘small-world’ networks. *nature*, 393(6684):440–442, 1998.
- [20] S. Zejnilovic, J. Gomes, and B. Sinopoli. Network observability and localization of the source of diffusion based on a subset of nodes. In *Proc. Annu. Allerton Conf. Communication, Control and Computing*, Monticello, IL, 2013.
- [21] K. Zhu and L. Ying. Information source detection in the SIR model: A sample path based approach. In *Proc. Information Theory and Applications Workshop (ITA)*, Feb. 2013.
- [22] K. Zhu and L. Ying. A robust information source estimator with sparse observations. In *Proc. IEEE Int. Conf. Computer Communications (INFOCOM)*, Toronto, Canada, April-May 2014.

- [23] K. Zhu and L. Ying. Source localization in networks: Trees and beyond. *arXiv:1510.01814*, 2015.
- [24] K. Zhu and L. Ying. Information source detection in networks: Possibility and impossibility results. In *Proc. IEEE Int. Conf. Computer Communications (INFOCOM)*, San Francisco, CA, 2016.
- [25] Kai Zhu, Zhen Chen, and Lei Ying. Locating the contagion source in networks with partial timestamps. *Data Mining and Knowledge Discovery*, 2015.

Appendices

A Proof of E_1 and E_2

Lemma 1. Assume the conditions in Theorem 2 hold, for any $\epsilon > 0$, we have

$$\Pr(E_1) \geq 1 - \epsilon,$$

and

$$\Pr(E_2|E_1) \geq 1 - \epsilon,$$

for sufficiently large n .

The proof follows directly from the proof of Lemma 3 and 4 in [23].

B Proof of E_3

Lemma 2. Assume the conditions in Theorem 2 hold, for any $\epsilon > 0$, we have

$$\Pr(E_3|E_1) \geq 1 - \epsilon,$$

for sufficiently large n .

Proof. Since all sources are within D hops from s_1 and the snapshot is taken at time t , all the infected nodes are within $t + D$ hops from s_1 . To prove the conclusion, we only need to show that any node on level $t + D + 1$ does not have more than $(1 - \delta)^3 \mu q \theta$ neighbors. Since all infected nodes are within level $t + D$, instead of considering the observed infected neighbors, we only need to show that any node on level $t + D + 1$ does not have more than $(1 - \delta)^3 \mu q \theta$ neighbors in level $t + D$. Therefore, in this proof, we consider a more restrictive event which is only a topological feature of the ER random and does not depend on the infection process.

Based on E_1 , there are at most $[(1 + \delta)\mu]^{t+D}$ nodes in level $t + D$ and at most $[(1 + \delta)\mu]^{t+D+1}$ nodes in level $t + D + 1$. For any node v in level $t + D + 1$, the neighbors of node v in $t + D$ are either the node which introduce node v into level $t + D + 1$ (i.e., the parent of v in the BFS tree) or the collision edges between node v and nodes in level $t + D$. The total number of possible collision edges depends on the order that the parent of node v is introduced to the BFS tree.

In general, if the parent of node v is the i th node, the number of possible neighbors on level $t + D$ follows $\text{Bi}([(1 + \delta)\mu]^{D+t} - i, \mu/n) + 1$. As a summary, for any node on level $t + D + 1$ the number of neighbors in level $t + D$ is stochastically upper bounded by $\text{Bi}([(1 + \delta)\mu]^{D+t}, \mu n) + 1$. Define

$$X \triangleq (1 - \delta)^3 \mu q \theta - 1.$$

Define δ' to be

$$\delta' \triangleq \frac{Xn}{[(1 + \delta)\mu]^t \mu} - 1.$$

Denote by N_v the number of neighbors in level $t + D$ for one node v on level $t + D + 1$.

$$\Pr(N_v \geq X + 1 | E_1) \leq \exp \left(-\frac{\delta'^2 [(1 + \delta)\mu]^{t+D} \mu/n}{2 + \delta'} \right) \quad (7)$$

$$= \exp \left(-\frac{\delta'}{2+\delta'} \times \delta' [(1+\delta)\mu]^{t+D} \mu/n \right) \quad (8)$$

$$= \exp \left(-\frac{\delta'}{2+\delta'} \times \left(\frac{Xn}{[(1+\delta)\mu]^{t+D}\mu} - 1 \right) [(1+\delta)\mu]^{t+D} \mu/n \right) \quad (9)$$

$$= \exp \left(-\frac{\delta'}{2+\delta'} \times \left(\frac{X}{[(1+\delta)\mu]^{t+D}\mu} [(1+\delta)\mu]^{t+D} \mu - [(1+\delta)\mu]^{t+D} \mu/n \right) \right) \quad (10)$$

$$= \exp \left(-\frac{\delta'}{2+\delta'} \times (X - [(1+\delta)\mu]^{t+D} \mu/n) \right) \quad (11)$$

$$= \exp \left(-\frac{\delta'}{2+\delta'} \times ((1-\delta)^3 \mu q \theta - 1 - [(1+\delta)\mu]^{t+D} \mu/n) \right) \quad (12)$$

$$= \exp \left(-\frac{\delta'\mu}{2+\delta'} ((1-\delta)^3 q \theta - 1/\mu - [(1+\delta)\mu]^{t+D}/n) \right) \quad (13)$$

$$\leq \exp \left(-\frac{\delta'(1-\delta)^3 q \theta \mu}{2(2+\delta')} \right) \quad (14)$$

$$\leq \exp \left(-\frac{(1-\delta)^3 q \theta}{6} \mu \right) \quad (15)$$

Since $t + D < \frac{\log n}{(1+\alpha)\log \mu}$, we have

$$t + D \leq \frac{\log n}{\log \mu} \times \frac{1 - \frac{\log \mu}{\log n}}{1 + \frac{\log(1+\delta)}{\log \mu}}. \quad (16)$$

Hence,

$$\frac{[(1+\delta)\mu]^{t+D}}{n} \leq \frac{1}{\mu} \quad (17)$$

Hence, Inequality (14) is based on Inequality (17) and Inequality (15) is based on $\delta' \geq 1$ for sufficiently large n . For any node in level $t + D + 1$, we have

$$\begin{aligned} & \Pr(\cap_v N_v < X + 1 | E_1) \\ &= 1 - \Pr(\cup_v N_v \geq X + 1 | E_1) \\ &\geq 1 - \sum_v \Pr(N_v \geq X + 1 | E_1) \\ &\geq 1 - [(1+\delta)\mu]^{t+D+1} \exp(-\Omega(\mu)) \\ &\geq 1 - \exp \left((t + D + 1) \log[(1+\delta)\mu] - \frac{(1-\delta)^3 q \theta}{6} \mu \right) \end{aligned}$$

Note we have $t + D \leq \frac{\log n}{(1+\alpha)\log \mu}$. Therefore,

$$\begin{aligned} & \Pr(\cap_v N_v < X + 1 | E_1) \\ &\geq 1 - \exp \left(((t + D) \log[(1+\delta)] + (t + D) \log \mu + \log[(1+\delta)\mu]) - \frac{(1-\delta)^3 q \theta}{6} \mu \right) \end{aligned}$$

$$\geq 1 - \exp \left(\frac{\log n}{(1+\alpha) \log \mu} \log[(1+\delta)] + \frac{\log n}{(1+\alpha)} + \log[(1+\delta)\mu] - \frac{(1-\delta)^3 q \theta}{6} \mu \right)$$

Let $C \leq \frac{6}{(1-\delta)^3(1+\alpha)}$ and since $\mu > \frac{1}{Cq\theta} \log n$, we have

$$\Pr(\cap_v N_v < X + 1) \geq 1 - \exp(-\Omega(\mu))$$

for sufficiently large n . By substituting X , we proved the lemma. \square

C Proof of E_4 and E_5

C.1 Neighboring structure of all sources

To prove E_4 and E_5 happen with a high probability, we first analyze the neighborhood of all sources in the ER random graph. In this section, we derived upper and lower bounds of the t neighborhood of all sources. Define \mathcal{L}_l^i the set of nodes from level 0 to level l of the BFS tree rooted in source s_i . In addition, define $\phi'_i(v)$ the number of offsprings of node v on the BFS tree rooted in source s_i . Define

$$E_1^i = \{\forall v \in \mathcal{L}_{t-1}^i, \phi'_i(v) \in ((1-\delta)\mu, (1+\delta)\mu)\}$$

Denote by the event

$$\tilde{E} = \cap_{i=2}^m E_1^i \cap E_1.$$

We have the following lemma.

Lemma 3. *If the conditions in Theorem 2 hold, for any $\epsilon > 0$,*

$$\Pr(\tilde{E}) \geq 1 - \epsilon$$

for sufficiently large n .

Proof. For each infection source s_i , follow the similar argument of Lemma 3 in [23], we have

$$\Pr(E_1^i) \geq \exp \left(-8 \exp \left(-\frac{\delta^2 \mu}{2+\delta} + (t-1) \log[(1+\delta)\mu] \right) \right) \geq 1 - \frac{8(\mu(1+\delta))^{t-1}}{\exp \left(\frac{\delta^2 \mu}{2+\delta} \right)}$$

Hence, we have

$$\Pr(\bar{E}_1^i) \leq \frac{8(\mu(1+\delta))^{t-1}}{\exp \left(\frac{\delta^2 \mu}{2+\delta} \right)}$$

Therefore, with an union bound, we have

$$\begin{aligned} \Pr(\cap_{i=2}^m E_1^i) &\geq 1 - \sum_{i=1}^m \Pr(\bar{E}_1^i) \\ &\geq 1 - \frac{8(m-1)(\mu(1+\delta))^{t-1}}{\exp \left(\frac{\delta^2 \mu}{2+\delta} \right)} \\ &\geq 1 - \exp \left(\log 8(m-1) + (t-1) \log(\mu(1+\delta)) - \frac{\delta^2 \mu}{2+\delta} \right) \end{aligned}$$

To make the probability larger than $1 - \epsilon$, we have

$$t \leq \frac{\log \frac{\epsilon}{8(m-1)} + \frac{\delta^2 \mu}{2+\delta}}{\log(\mu(1+\delta))} + 1$$

Again, we have $t + D < \frac{\log n}{(1+\alpha) \log \mu}$ and $\mu > \frac{1}{Cq\theta} \log n > 3 \log n$ which guarantees the probability goes to 1 asymptotically.

Note the events $\cap_{i=2}^m E_1^i$ do not contain the neighborhood of source s_1 . Based on Lemma 1, with a union bound, it is straightforward to show that

$$\Pr(\cap_{i=2}^m E_1^i \cap E_1) > 1 - \epsilon.$$

for sufficiently large n . Hence the lemma is proved. □

C.2 Proof of E_4 and E_5

Lemma 4. *If the conditions in Theorem 2 hold, for any $\epsilon > 0$,*

$$\Pr(E_4, E_5) \geq 1 - \epsilon$$

for sufficiently large n .

Proof. We still consider the BFS tree rooted at source s_1 and we can rewrite the events E_4 and E_5 in a combined fashion

$$E_4 \cap E_5 = \{\forall v \in \cup_{i=0}^{t-1} \mathcal{Z}_i, \psi'(v) \geq (1-\delta)^2 \mu q, \psi''(v) \geq (1-\delta)^3 \mu q \theta\}.$$

Define

$$\mathcal{F}_i = \{\mathcal{Z}_i | \forall v \in \mathcal{Z}_i, \psi'(v) \geq (1-\delta)^2 \mu q, \psi''(v) \geq (1-\delta)^3 \mu q \theta\}$$

Then, we have

$$\begin{aligned} & \Pr(E_4, E_5) \\ & \geq \Pr(E_4, E_5 | \tilde{E}) \Pr(\tilde{E}) \end{aligned}$$

$$\begin{aligned} & \Pr(E_4, E_5 | \tilde{E}) \\ & = \Pr(\forall v \in \cup_{i=0}^{t-1} \mathcal{Z}_i, \psi'(v) \geq (1-\delta)^2 \mu q, \psi''(v) \geq (1-\delta)^3 \mu q \theta | \tilde{E}) \\ & = \sum_{\mathcal{Z}_1 \in \mathcal{F}_1} \cdots \sum_{\mathcal{Z}_{t-1} \in \mathcal{F}_{t-1}} \Pr(\forall v \in \mathcal{Z}_{t-1}, \psi'(v) > (1-\delta)^2 \mu q, \psi''(v) \geq (1-\delta)^3 \mu q \theta \\ & \quad | \mathcal{Z}_{t-1}, \mathcal{Z}_{t-2}, \dots, \mathcal{Z}_1, \tilde{E}) \Pr(\mathcal{Z}_{t-1}, \mathcal{Z}_{t-2}, \dots, \mathcal{Z}_1 | \tilde{E}) \end{aligned}$$

We have

$$\Pr(\forall v \in \mathcal{Z}_{t-1}, \psi'(v) > (1-\delta)^2 \mu q | \mathcal{Z}_{t-1}, \mathcal{Z}_{t-2}, \dots, \mathcal{Z}_1, \tilde{E})$$

$$\geq 1 - \sum_{v \in \mathcal{Z}_{t-1}} \Pr(\psi'(v) \leq (1-\delta)^2 \mu q, \psi''(v) \geq (1-\delta)^3 \mu q \theta | \mathcal{Z}_{t-1}, \mathcal{Z}_{t-2}, \dots, \mathcal{Z}_1, \tilde{E})$$

Note conditioned on $\mathcal{Z}_{t-1}, \mathcal{Z}_{t-2}, \dots, \mathcal{Z}_1$, consider an offspring u of node v on the BFS tree rooted at source s_1 . Node u has two possible states: infected or susceptible. If u is not infected, v will infect node u with probability q in the next time slot. On the other hand, if u is infected, it counts as an infected offspring of node v deterministically. Therefore, $\psi'(v)$ is stochastically lower bounded by binomial distribution $B((1-\delta)\mu, q)$. Therefore, with Chernoff bound in [23], we have

$$\Pr(\psi'(v) \leq (1-\delta)^2 \mu q | \mathcal{Z}_{t-1}, \mathcal{Z}_{t-2}, \dots, \mathcal{Z}_1, \tilde{E}) \leq \exp\left(-\frac{\delta^2(1-\delta)\mu q}{2}\right)$$

Each infected nodes are observed with probability θ independently. Conditioned on $\psi'(v) \geq (1-\delta)^2 \mu q$, $\psi''(v)$ is stochastically lower bounded by $B((1-\delta)^2 \mu q, \theta)$. Therefore,

$$\Pr(\psi''(v) \leq (1-\delta)^3 \mu q \theta | \psi'(v) \geq (1-\delta)^2 \mu q, \mathcal{Z}_{t-1}, \mathcal{Z}_{t-2}, \dots, \mathcal{Z}_1, \tilde{E}) \leq \exp\left(-\frac{\delta^2(1-\delta)^2 \mu q \theta}{2}\right)$$

Therefore, we have

$$\begin{aligned} & \Pr(\psi''(v) \geq (1-\delta)^3 \mu q \theta, \psi'(v) \geq (1-\delta)^2 \mu q | \mathcal{Z}_{t-1}, \mathcal{Z}_{t-2}, \dots, \mathcal{Z}_1, \tilde{E}) \\ &= \left(1 - \exp\left(-\frac{\delta^2(1-\delta)\mu q}{2}\right)\right) \left(1 - \exp\left(-\frac{\delta^2(1-\delta)^2 \mu q \theta}{2}\right)\right) \\ &\geq 1 - \exp\left(-\frac{\delta^2(1-\delta)\mu q}{2}\right) - \exp\left(-\frac{\delta^2(1-\delta)^2 \mu q \theta}{2}\right) \\ &\geq 1 - 2 \exp\left(-\frac{\delta^2(1-\delta)^2 \mu q \theta}{2}\right) \end{aligned}$$

Again with union bound, we have

$$\begin{aligned} & \Pr(\forall v \in \mathcal{Z}_{t-1}, \psi'(v) > (1-\delta)^2 \mu q, \psi''(v) \geq (1-\delta)^3 \mu q \theta | \mathcal{Z}_{t-1}, \mathcal{Z}_{t-2}, \dots, \mathcal{Z}_1, \tilde{E}) \\ &\geq 1 - 2|\mathcal{Z}_{t-1}| \exp\left(-\frac{\delta^2(1-\delta)^2 \mu q \theta}{2}\right) \end{aligned}$$

Note, based on event \tilde{E} , we have

$$|\mathcal{Z}_{t-1}| \leq m \sum_{i=0}^{t-1} [(1+\delta)\mu]^i \leq 2m[(1+\delta)\mu]^{t-1}$$

Hence, we have

$$\begin{aligned} & \Pr(\forall v \in \mathcal{Z}_{t-1}, \psi'(v) > (1-\delta)^2 \mu q, \psi''(v) \geq (1-\delta)^3 \mu q \theta | \mathcal{Z}_{t-1}, \mathcal{Z}_{t-2}, \dots, \mathcal{Z}_1, \tilde{E}) \\ &\geq 1 - 4m[(1+\delta)\mu]^{t-1} \exp\left(-\frac{\delta^2(1-\delta)^2 \mu q \theta}{2}\right) \end{aligned}$$

Therefore, we have

$$\Pr(E_4, E_5 | \tilde{E})$$

$$\begin{aligned}
&= \sum_{\mathcal{Z}_1 \in \mathcal{F}_1} \cdots \sum_{\mathcal{Z}_{t-1} \in \mathcal{F}_{t-1}} \Pr(\forall v \in \mathcal{Z}_{t-1}, \psi'(v) > (1-\delta)^2 \mu q, \psi''(v) \geq (1-\delta)^3 \mu q \theta \\
&\quad | \mathcal{Z}_{t-1}, \mathcal{Z}_{t-2}, \dots, \mathcal{Z}_1, \tilde{E}) \Pr(\mathcal{Z}_{t-1}, \mathcal{Z}_{t-2}, \dots, \mathcal{Z}_1 | \tilde{E}) \\
&\geq \sum_{\mathcal{Z}_1 \in \mathcal{F}_1} \cdots \sum_{\mathcal{Z}_{t-1} \in \mathcal{F}_{t-1}} \left(1 - 4m[(1+\delta)\mu]^{t-1} \exp\left(-\frac{\delta^2(1-\delta)^2 \mu q \theta}{2}\right) \right) \\
&\quad \times \Pr(\mathcal{Z}_{t-1}, \mathcal{Z}_{t-2}, \dots, \mathcal{Z}_1 | \tilde{E}) \\
&= \left(1 - 4m[(1+\delta)\mu]^{t-1} \exp\left(-\frac{\delta^2(1-\delta)^2 \mu q \theta}{2}\right) \right) \\
&\quad \times \sum_{\mathcal{Z}_1 \in \mathcal{F}_1} \cdots \sum_{\mathcal{Z}_{t-2} \in \mathcal{F}_{t-2}} \Pr(\forall v \in \mathcal{Z}_{t-2}, \psi'(v) > (1-\delta)^2 \mu q, \psi''(v) \geq (1-\delta)^3 \mu q \theta \\
&\quad | \mathcal{Z}_{t-2}, \mathcal{Z}_{t-3}, \dots, \mathcal{Z}_1, \tilde{E}) \Pr(\mathcal{Z}_{t-2}, \mathcal{Z}_{t-3}, \dots, \mathcal{Z}_1 | \tilde{E})
\end{aligned}$$

Then, iteratively apply the similar arguments, we obtain,

$$\Pr(E_4, E_5 | \tilde{E}) \tag{18}$$

$$\geq \prod_{i=2}^t \left(1 - 4m[(1+\delta)\mu]^{i-1} \exp\left(-\frac{\delta^2(1-\delta)^2 \mu q \theta}{2}\right) \right) \tag{19}$$

$$= \prod_{i=2}^t \left(1 - \exp\left(\log 4m + (i-1) \log[(1+\delta)\mu] - \frac{\delta^2(1-\delta)^2 \mu q \theta}{2}\right) \right) \tag{20}$$

$$\geq \prod_{i=2}^t \exp\left(-2 \exp\left(\log 4m + (i-1) \log[(1+\delta)\mu] - \frac{\delta^2(1-\delta)^2 \mu q \theta}{2}\right)\right) \tag{21}$$

$$= \exp\left(-2 \sum_{i=2}^t \exp\left(\log 4m + (i-1) \log[(1+\delta)\mu] - \frac{\delta^2(1-\delta)^2 \mu q \theta}{2}\right)\right) \tag{22}$$

$$= \exp\left(-2 \exp\left(\log 4m - \frac{\delta^2(1-\delta)^2 \mu q \theta}{2}\right) \sum_{i=1}^{t-1} \exp(i \log[(1+\delta)\mu])\right) \tag{23}$$

$$\geq \exp\left(-4 \exp\left(\log 4m - \frac{\delta^2(1-\delta)^2 \mu q \theta}{2} + (t-1) \log[(1+\delta)\mu]\right)\right) \tag{24}$$

To make the probability greater than $1 - \epsilon$, we need,

$$t \leq 1 + \frac{\log \log(1-\epsilon)^{-1/4} - \log 2m + \frac{\delta^2(1-\delta)^2 \mu q \theta}{2}}{\log((1+\delta)\mu)} \tag{25}$$

Since we have $t + D < \frac{\log n}{(1+\alpha) \log \mu}$ and $\mu > \frac{1}{Cq\theta} \log n > \frac{2}{\delta^2(1-\delta)^2 q \theta} \log n$, Inequality 25 is satisfied when n is large enough.

Therefore, based on Lemma 3 and Inequality 24, we showed that

$$\Pr(E_4, E_5) \geq \Pr(E_4, E_5 | \tilde{E}) \Pr(\tilde{E}) \geq 1 - \epsilon$$

for any $\epsilon > 0$ when n is sufficiently large. \square

D Proof of E_6

Lemma 5. *If the conditions in Theorem 2 hold, for any $\epsilon > 0$,*

$$\Pr(E_6|E_1) \geq 1 - \epsilon$$

for sufficiently large n .

Proof. Define

$$E_7 = \{\tilde{Z}_1^1 \geq (1 - \delta)^2 \mu q\} \cap \{\forall v \in \tilde{Z}_1^1, \cap_{i=2}^t \tilde{Z}_i^i(v) \geq (1 - \delta)^2 \mu q \tilde{Z}_{i-1}^{i-1}(v)\}$$

Following the similar arguments in Lemma 5 in [23], we have

$$\Pr(E_7|E_1) \geq 1 - \epsilon.$$

Note, for each node $v \in \tilde{Z}_1^1$, based on event E_7 , we have $\tilde{Z}_t^t(v) \geq [(1 - \delta)^2 \mu q]^{t-1}$. Recall each infected node report its status independently. Therefore, $\tilde{Z}_t^t(v)$ is stochastically lower bounded by $\text{Bi}([(1 - \delta)^2 \mu q]^{t-1}, \theta)$. By Chernoff bound, we have

$$\Pr(\tilde{Z}_t^t(v) \geq [(1 - \delta)^2 \mu q]^{t-1} (1 - \delta) \theta | E_7, E_1) \leq \exp\left(-\frac{\delta^2 [(1 - \delta)^2 \mu q]^{t-1} \theta}{2}\right)$$

Note $\tilde{Z}_1^1 \leq (1 + \delta) \mu$ based on event E_1 , with a union bound, we have

$$\begin{aligned} & \Pr(\forall v \in \tilde{Z}_1^1, \tilde{Z}_t^t(v) \geq [(1 - \delta)^2 \mu q]^{t-1} (1 - \delta) \theta | E_7, E_1) \\ & \geq 1 - (1 + \delta) \mu \exp\left(-\frac{\delta^2 [(1 - \delta)^2 \mu q]^{t-1} \theta}{2}\right) \\ & \geq 1 - \exp\left(\log[(1 + \delta) \mu] - \frac{\delta^2 [(1 - \delta)^2 \mu q]^{t-1} \theta}{2}\right) \\ & \geq 1 - \epsilon, \end{aligned}$$

for sufficiently large n since $\mu > \frac{1}{Cq\theta} \log n$.

Therefore, we have

$$\Pr(E_6|E_1) \geq \Pr(E_6|E_7, E_1) \Pr(E_7|E_1) \geq 1 - \epsilon.$$

The lemma is proved. □